Анна Вайнберг Аллен

Графы для анализа структурных соотношений между переменными и их приложение к изучению российских регионов (Часть 1)

Работа посвящена проблемам исследования структуры переменных. Эта проблематика имеет важное значение для многих статистических приложений, и, в частности, в межстрановом и межрегиональном сравнительном анализе. Статья состоит из двух частей, включающих (1) обзор двух методов исследования структуры переменных: краткого резюме по древообразным структурам зависимостей и более подробного представления графовых моделей с особым упором на метод выбора ковариаций Демпстера (часть 1), (2) описания предлагаемой автором модификации этого метода и (3) его применения к практическому исследованию и сравнению российских регионов (часть 2).

> Сколько нужно гипотетических переменных, или факторов, и каких, чтобы по возможности более точно воспроизвести и объяснить наблюдаемые взаимосвязи между переменными? Каким образом мы можем провести свертку большого количества данных с минимальной потерей информации, используя возможную более простую концепцию?
>
> *Иберла* [Iberla (1980)]

Избыток переменных — проблема многих сравнительных исследований. Как правило, трудно принять решение о полезности и уместности отдельной специфической переменной. В этой ситуации часто применяется экспертное агрегирование переменных в новые, более общие факторы. Но при таком подходе возникает хорошо известная проблема произвольности, и поэтому обоснован интерес к исследованию структуры многомерной случайной переменной.

Этот интерес, или лучше сказать потребность, обычная в общественных и точных науках, породила ряд методов и алгоритмов, основанных на теории графов.

В первой части статьи описываются два из возможных методов: кратко излагается метод, базирующийся на понятии «деревья зависимостей» (хорошо представленный в русскоязычной литературе), и, более подробно, алгоритм выбора ковариаций Демпстера, гораздо менее известный.

Последний метод относится к обширной и быстро развивающейся области графовых моделей. Предлагается также краткое введение в графовые модели, при этом предлагается алгоритм выбора ковариаций относительно других методов, используемых в этой области.

Во второй части статьи рассматривается модификация алгоритма Демпстера, основанная на его комбинации с деревьями зависимостей, и описывается его практическое применение к сравнительному анализу положения и развития российских регионов.

1. Древообразные структуры зависимостей

1.1. Общее представление

Традиционной формой описания зависимостей между p элементами многомерного нормально распределенного вектора является корреляционная матрица, содержащая полную информацию о взаимосвязях между координатами. В общем случае эта матрица зависит от $p(p-1)/2$ параметров, и в дальнейших вычислениях или интерпретациях нелегко обработать такое большое количество параметров.

Чоу и Лью [Chow and Liu (1966)], [Chow and Liu (1968)], [Chow (1970)] на основе теории графов ввели понятие древообразной структуры зависимостей, предложив новый класс распределений с более экономным описанием корреляций, используя только $p-1$ параметров. Дальнейшие разработки этой темы для пространств высокой размерности были выполнены Заруцким [Заруцкий (1978)] и [Заруцкий (1980)]. Практический пример использования приведен в [Прохорскас, Жюжнис и др. (1976)]. Наше представление древообразных структур зависимостей заимствовано у Айвазяна, Енюкова и Мешалкина [Айвазян, Енюков и др. (1985)], а сама идея древообразных структур зависимостей происходит из цепей Маркова.

Определение 1 (Цепь Маркова)[1]. *Последовательность X_i случайных непрерывных переменных, обладающих тем свойством, что, при данном представлении, будущее условно независимо от прошлого, называется цепью Маркова. Другими словами, для всех $i \geq 2$*

$$f_{X_i}(x_i | X_{i-1} = x_{i-1}) = f_{X_i}(x_i | X_{i-1} = x_{i-1}, \ldots, X_1 = x_1).$$

Определение 2 (Древообразная структура зависимостей). *Случайный p-мерный вектор X можно представить в виде древообразной структуры зависимостей T, если существует по крайней мере одна перестановка $\alpha(1, \ldots, p) = (\alpha_{(1)}, \alpha_{(2)}, \ldots, \alpha_{(p)})$ координат этого вектора, удовлетворяющая условию: для каждого $\alpha_{(i)}$ существует такое число $j(\alpha_{(i)}) \in 0, \alpha_{(1)}, \ldots, \alpha_{(i-1)}$, что*

$$f_{X_{a_{(i)}}}(x_{a_{(i)}} | X_{j(a_{(1)})} = x_{j(a_{(1)})}) = f_{X_{a_{(i)}}}(x_{a_{(i)}} | X_{a_{(1)}} = x_{a_{(1)}}, \ldots, X_{a_{(i-1)}} = x_{a_{(i-1)}}). \quad (1)$$

Здесь $j = 0$ соответствует фиктивной координате $x^0 \equiv 1$ и $j(\alpha_{(1)}) = 0$.

Проиллюстрируем это описание с помощью примера Демпстера [Dempster (1972)], представленного на рис. 1 и соответствующего корреляционной матрице S с числом переменных $p = 6$ и числом наблюдений $n = 72$:

$$S = \begin{pmatrix} 1{,}0000 & 0{,}3966 & 0{,}3688 & 0{,}1764 & -0{,}4632 & 0{,}2939 \\ & 1{,}0000 & 0{,}0232 & -0{,}0854 & 0{,}0193 & 0{,}2191 \\ & & 1{,}0000 & 0{,}0494 & -0{,}1350 & -0{,}2376 \\ & & & 1{,}0000 & -0{,}4671 & 0{,}1135 \\ & & & & 1{,}0000 & -0{,}3656 \\ & & & & & 1{,}0000 \end{pmatrix}. \quad (2)$$

[1] В данном случае мы предполагаем наличие плотности распределения. Общее определение может быть найдено в книге [Ширяев (2004)], стр. 788.

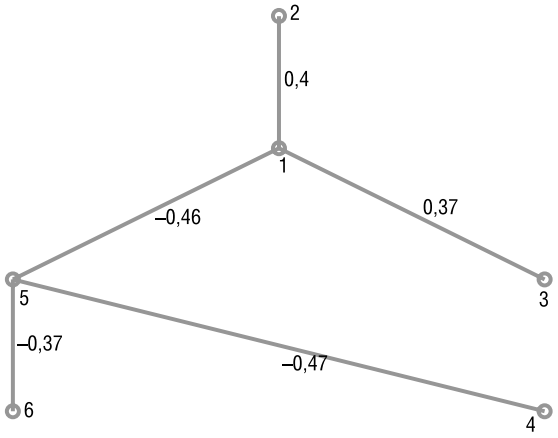

Рис. 1. Граф для примера Демпстера

Для шестимерного случайного вектора $X = (X_1, X_2, X_3, X_4, X_5, X_6)$ здесь существует перестановка

$$\alpha = (\alpha_{(1)}, \alpha_{(2)}, \alpha_{(3)}, \alpha_{(4)}, \alpha_{(5)}, \alpha_{(6)}) = (2, 1, 3, 5, 4, 6),$$

где $j(\alpha_{(1)}) = j(2) = 0$, $j(\alpha_{(2)}) = j(1) = 2$, $j(\alpha_{(3)}) = j(3) = 1$, $j(\alpha_{(4)}) = j(5) = 1$, $j(\alpha_{(5)}) = j(4) = 5$ и $j(\alpha_{(6)}) = j(6) = 5$.

Для нормального распределения эта древообразная структура зависимостей единственна [Chow (1970)] и [Заруцкий (1978)]. Кроме того, нормальное распределение со свойством древообразной структуры зависимостей имеет простой вид обратной ковариационной матрицы Σ^{-1}, где верхняя треугольная часть матрицы Σ^{-1} содержит не более $p-1$ ненулевых элементов. Вспомним определение частичной независимости.

Определение 3 (Частичная независимость). *Две переменные частично независимы (при фиксированных всех остальных переменных) тогда и только тогда, когда соответствующий элемент в обратной ковариационной матрице равен нулю.*

Таким образом, в нашем случае, за исключением $p-1$ пар переменных, все другие переменные частично независимы.

В этом случае справедлива следующая теорема:

Теорема 1 (Свойство цепи, [Chow (1970)]). *Для p-мерного нормально распределенного вектора X с древообразной структурой зависимостей T для всех* $1 \le i < j \le p$

$$\rho_{ij} = \prod_{(k,l) \in M(i,j)} \rho_{kl}, \tag{3}$$

где ρ_{ij} *— коэффициент корреляции*[2] *между векторными координатами* X_i *и* X_j, *а* $M(i, j)$ *— простая цепь*[3], соединяющая вершины *i и j графа T*.

[2] Мы обозначаем через ρ_{ij} «теоретические» коэффициенты корреляции, а через r_{ij} — выборочные коэффициенты корреляции.

[3] Напомним:

Определение 4 (Простая цепь). *Конечная непустая последовательность* $M = (x_1, x_2), (x_2, x_3), \ldots, (x_m, x_{m+1})$ *ребер графа называется простой цепью, соединяющей вершины* x_1 *и* x_{m+1}, *если все вершины* $x_1, \ldots, x_{m+1}$ *различны. Если* $x_1 = x_{m+1}$, *то мы имеем простой цикл.*

Далее нам потребуются следующие два определения:

Определение 5 (Вес связи). *Назовем весом w_{ij} связи (i, j) абсолютное значение ρ_{ij}.*

Определение 6 (Вес графа). *Вес графа — это сумма весов его ребер.*

Следующая ключевая теорема дает возможность практического построения графа.

Теорема 2 (Чоу (Chow)). *Рассмотрим невырожденный нормальный вектор с древообразной структурой T. Вес этого дерева зависимостей строго больше, чем вес любого другого дерева, отличающегося, по крайней мере, одним ребром с ненулевым весом.*

Доказательства теорем 1 и 2 приведены в [Айвазян, Енюков и др. (1985)]. Так, в случае известной корреляционной матрицы $\Sigma = \rho_{ij}$ проблема поиска древообразной структуры зависимостей *T* эквивалентна поиску дерева с максимальным весом. Эта проблема решается с помощью алгоритма Крускала [Kruskal (1956)] поиска максимального связывающего дерева. Сложность алгоритма Крускала оценивается как $O(p^2 \log(p))$ [Cormen, Leiberson, et al. (1990)].

Древообразные структуры зависимостей предоставляют полезную классификацию переменных и позволяют осуществлять неявный контроль качества данных. В частности, те переменные в структуре, которые занимают необычное, не объясняемое теорией или здравым смыслом место, требуют проведения дополнительных исследований для выяснения причин таких отклонений.

Древообразные структуры зависимостей были разработаны для бинарных, а также нормально распределенных данных. В дальнейшем предполагается, что наши данные распределены нормально, и что справедлива сама гипотеза присутствия структуры древообразной зависимости в рассматриваемых данных.

В следующем разделе вводятся критерии качества представления.

1.2. Критерии качества представления

Обозначим через r_{ij} выборочный коэффициент корреляции, а через $\hat{r}_{ij}$ его оценку на основе древообразной структуры зависимостей. Если ребро (i, j) принадлежит дереву с максимальным весом (корреляцией), то

$$\hat{r}_{ij} = r_{ij},$$

в противном случае $\hat{r}_{ij}$ вычисляется, опираясь на выражение (3).

Хотим оценить, насколько хорошо древообразная структура зависимостей аппроксимирует всю корреляционную матрицу и корреляции каждой переменной *i* со всеми другими переменными. Задаем три критерия качества представления.

Критерий 1: качество представления переменной

Для заданной переменной *i*:

Если $$a_i = \sum_{j \neq i} |r_{ij} - \hat{r}_{ij}| \text{ и } b_i = \sum_{j \neq i} |r_{ij}|,$$

тогда $$c_i = 1 - \frac{a_i}{b_i} \qquad (4)$$

определяет качество представления переменной *i*.

Значение критерия c_i равно 1, если и только если $\hat{r}_{ij} = r_{ij}$ для всех j, и, таким образом, значение c_i, близкое к 1, указывает на хорошее качество представления. Критерий c_i достигает своего наихудшего значения, равного –1, если и только если $\hat{r}_{ij} = -r_{ij}$ для всех j.

Критерий 2: качество представления переменной (робастное)

Другая возможность состоит в том, чтобы определить для данной переменной i

$$c_i^* = 1 - \frac{a_i}{\max_i(\delta, b_i)}, \tag{5}$$

где δ — порог индифферентности.

Вводим этот порог, чтобы ограничить влияние слишком низких корреляций. Используем $\delta = (p-1)\overline{|r|}$ или $\delta = \frac{2}{3}(p-1)\overline{|r|}$, где $\overline{|r|}$ — среднее значение по всем $|r_{ij}|$, а p — число переменных. Можно использовать порог δ, равный $(p-1)$, умноженный на максимальное значение $|r|$, для которого гипотеза $\rho = 0$[4] для выборки соответствующего объема не отклоняется.

Результаты для критериев 1 и 2 близки, но определение (5) более робастно и предотвращает получение экстремальных значений.

Критерий 3 (глобальный): качество представления дерева

Если $$A = \sum_{i<j} |r_{ij} - \bar{r}_{ij}| \text{ и } B = \sum_{i<j} |r_{ij}|,$$

тогда $$C = 1 - \frac{A}{B} \tag{6}$$

отражает качество представления дерева. Значения C также изменяются между –1 и 1. Значение критерия C, близкое к нулю, указывает на хорошее качество представления. По аналогии мы можем определить C в виде:

$$C = 1 - \frac{A}{\max(\delta, B)}, \tag{7}$$

где δ равно $p(p-1)$, умноженному на максимальное значение $|r|$, для которого гипотеза $\rho = 0$ для выборки соответствующего объема не отклоняется.

В качестве примера рассмотрим дерево зависимостей, построенное для российских областей в 1994 году и представленное на рис. 2. Применение деревьев зависимостей требует, чтобы переменные были распределены по нормальному закону, в данном примере переменные преобразованы так, чтобы получить распределения, близкие к нормальному (подробности в Приложении D2 диссертации автора [Weinberg (2007)]).

Качество представления корреляционной матрицы древообразной структурой зависимостей, представленной на рис. 2, имеет значение 0,43, что является достаточно низкой цифрой. Качество зависит от двух факторов: числа связей и абсолютных значений коэффициентов корреляции, выбранных для построения древообразной структуры.

Введем следующее определение, которое понадобится в дальнейшем изложении.

[4] Обозначаем через ρ теоретическое значение r.

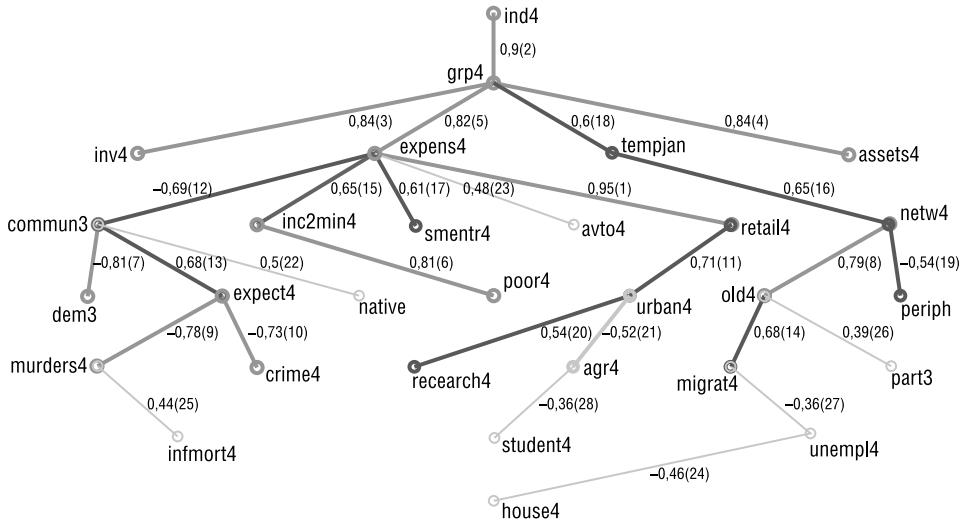

Рис. 2. Дерево зависимостей для российских областей в 1994 году. На ребрах указаны значения коэффициентов корреляции, числа в круглых скобках показывают порядок присоединения ребер. Например, 0,95(1) показывает, что ребро (expens4, retail4) с коэффициентом корреляции, равным 0,95, было выбрано первым

Определение 7 (Степень переменной). *Степенью переменной, т. е. вершины, является число исходящих ребер или, что эквивалентно, число смежных вершин.*

В табл. 1 приводятся значения критериев c_i качества представления переменных в этом примере. Качество представления зависит от двух факторов: числа связей и абсолютных значений коэффициентов корреляции, выбранных для построения дерева. В общем, для переменных с более высокими степенями качество представления лучше: значения c_i являются большими.

Таблица 1

Качество представления переменных в древообразной структуре зависимостей для российских регионов в 1994 году

Переменная	Степень	Качество	Переменная	Степень	Качество
expends4	6	0,82	migrat4	2	0,52
inv4	1	0,80	murders4	1	0,48
commun3	4	0,80	tempjan	2	0,47
grp4	5	0,78	**netw4**	3	0,43
retail4	2	0,76	inc2min4	2	0,41
dem3	1	0,74	poor4	1	0,39
assets4	1	0,69	student4	1	0,38
urban4	3	0,66	patr3	1	0,28
ind4	1	0,65	agr4	2	0,28
expect4	3	0,63	house4	1	0,27
crime4	1	0,59	unempl4	2	0,23
native	1	0,58	research4	1	0,19
avto4	1	0,58	infmort4	1	0,19
smentr4	1	0,57	periph	1	0,16
old4	3	0,54			

Однако, например, переменная инвестиции на душу населения (inv4) имеет только одну связь, а ее качество представления равно 0,80, что обусловлено высоким коэффициентом корреляции (0,9) переменной инвестиций с узловой переменной ВРП (GRP) (grp4). В то же самое время переменная плотности транспортной сети (netw4) с тремя связями имеет относительно низкое качество представления, равное 0,43. Это связано с тем, что коэффициент корреляции (0,65) между плотностью транспортной сети и температурой января (tempjan), приводящий к основному большому массиву данных и, таким образом, участвующий в большинстве цепей, имеет довольно низкое значение для такой ключевой переменной, высоко коррелированной с другими переменными. (Назовем ключевой переменной переменную с большим количеством связей, но строгое определение этому понятию не даем).

1.3. Интерпретация результатов

Рассмотрим структуру переменных нашей древообразной структуры зависимостей, начиная с ключевых переменных:

Переменная **ВРП на душу населения (grp4)** связана с переменными:

1) инвестиции на душу населения (inv4);
2) промышленное производство на душу населения (ind4);
3) розничная торговля на душу населения (retail4);
4) температура января (tempjan).

Таким образом, в дополнение к традиционной экономической значимости ВРП видим устойчивую географическую компоненту, обусловленную северным положением и вариацией ценового уровня по всей стране.

Расходы на душу населения (expens4). Эта переменная является другой важной экономической переменной, которая уже по построению агрегирует расходы на потребление. Эта переменная традиционно связана с отношением дохода к прожиточному минимуму (inc2min4), числом малых предприятий на душу населения (smentr4), числом автомобилей на душу населения (avto4) и оборотом розничной торговли на душу населения (retail4). Как видно, в 1994 году эта переменная также имела высокую отрицательную корреляцию с прокоммунистическим голосованием на выборах 1993 года (commun3). Области с более низкими уровнями благосостояния были менее благосклонны к либеральным реформам и переменам.

Во второй части статьи продолжим описание переменных, анализируя результаты, полученные при помощи графовых моделей. Увидим, что большинство связей (ребер графа) в древообразных структурах зависимостей и соответствующих графовых моделях совпадают. В целом, древообразная модель зависимостей приводит к удовлетворительным результатам, но в рамках этой модели на структуру переменных налагаются слишком строгие ограничения, в частности не допускаются циклы и изолированные переменные. Кроме того, в конце построения древообразной структуры часто добавляются недостаточно информативные ребра в ущерб более информативным. На рис. 2 порядок присоединения ребер указан в скобках, ребра с более высокими корреляциями, а значит, как правило, наиболее важные, согласно алгоритму добавляются первыми.

2. Графовые модели

Графовое моделирование — это вид многомерного анализа, в котором для представления моделей применяются графы...
Эдвардс [Edwards (2000)]

2.1. Введение в графовые модели

Согласно Лауритцену [Lauritzen (1996)] корни графовых моделей можно проследить в анализе траекторий и генетике [Wright (1921)], статистической механике [Gibbs (1902)] и анализе таблиц сопряженности [Bartlett (1935)]. Сначала графовые модели разрабатывались для построения моделей в каждой области научных интересов отдельно. Реальный интерес к графовым моделям появился в середине восьмидесятых [Borgelt, Kruse (2002)]. Структура условных отношений зависимости и независимости между переменными представлялась в виде сети или графа (отсюда названия — графовые модели и сети выводов (inference networks)) и часто называлась графом условной независимости. В таком графе каждая вершина представляет переменную, а каждое ребро — прямую зависимость между двумя переменными.

Такой граф оказывается не только полезным инструментом для представления содержания модели, но также облегчает анализ в областях с высокой размерностью, поскольку его применение позволяет свести анализ к подпространствам с более низкой размерностью.

Графовые модели часто применяются для получения выводов в экспертных системах и системах, обеспечивающих принятие решения [Castillo, Gutierrez, et al. (1997)]; в этом контексте они называются сетями выводов. Вероятностные графовые модели включают в себя: 1) *байесовские сети*, основанные на ориентированных графах условных независимостей [Jentzen (1996)], и 2) *марковские сети*, основанные на неориентированных графах [Lauritzen (1996)].

При обращении к проблемам выявления экспертных знаний (путем анализа внутренних связей данных) и «информационной проходке» (способ анализа информации в данных: выявление трендов и т. п.), графовые модели имеют определенные преимущества, что привело к возрастанию их популярности в последние годы. В частности, сетевое (графовое) представление обеспечивает понятное качественное описание (в виде сетевой структуры) и количественное описание (в виде соответствующих функций распределения) анализируемой области, так что с помощью полученных результатов исследования можно оценить достоверность интуиции экспертов [Borgelt, Kruse (2002)].

В статье рассмотрим традиционное приложение графовых моделей, т. е. построение модели в конкретной области исследования. В обширной области графового моделирования классическими являются три типа моделей:

1. *Логарифмически линейные модели для дискретных данных,* когда попарно условная независимость эквивалентна нулевым взаимодействиям между двумя факторами. Применение графов и графовых моделей совместно с известными и широко используемыми логарифмически линейными моделями началось с [Darroch, Lauritzen, et al.]; изложение с прикладной ориентацией представлено в [Edwards, Kreiner (1983)]; две другие важные ссылки [Whittaker (1990)] и [Lauritzen (1996)].

2. *Графовые гауссовские модели для непрерывных данных*. В этом случае попарно условная независимость соответствует нулевым частным корреляциям. После работы Демпстера [Dempster (1972)] графические гауссовские модели чаще называют моделями *выбора ковариаций*. Обзор этой темы можно найти в работе [Whittaker (1990)].

3. *Смешанные модели для смешанных непрерывных и дискретных данных.* Для смешанных непрерывных и дискретных данных модели ориентированных графов, а также модели неориентированных графов (последние также называются графовыми моделями взаимодействий) впервые описаны в статье [Lauritzen, Wermuth (1989)]. Объединяя логарифмически линейные модели для дискретных переменных с графовыми гауссовскими моделями для непрерывных переменных, Эдвардс [Edwards (1990)] обобщает класс неориентированных моделей на более широкий класс иерархических моделей.

Резюме этих трех типов моделей следует изложению монографии [Edwards (2000)], кроме того приводится описание программы MIM для реализации этих методов.

В статье рассматривается случай непрерывных переменных, и поэтому ограничим исследование вторым случаем графовых гауссовских моделей. Предположим, что $X = (X_1, \ldots, X_p)'$ — p-мерная случайная переменная, подчиняющаяся многомерному нормальному распределению со средним $\mu = (\mu_1, \ldots, \mu_p)'$ и с ковариационной матрицей $\Sigma = \{\sigma_{ij}\}$, и что существует обратная ковариационная матрица $\Sigma^{-1} = \{\sigma^{ij}\}$.

Особенно интересно исследовать значения элементов обратной ковариационной матрицы в случае, когда две переменные условно независимы (независимы при фиксированных остальных переменных). Графовые гауссовские модели определяются значениями элементов обратной ковариационной матрицы. Две переменные условно независимы тогда и только тогда, когда значение соответствующего элемента обратной ковариационной матрицы равно нулю.

Графовые методы моделирования позволяют строить графы со всеми узлами (переменными) и с такими ребрами, для которых соответствующие элементы в обратной корреляционной матрице, а следовательно, и соответствующие частные коэффициенты корреляции не равны нулю. Главная цель состоит в поиске множества частных коэффициентов корреляции и его графового представления с тем, чтобы получить выводы об анализируемом явлении.

Предлагаемый метод основан на аппроксимации истинной корреляционной матрицы корреляционной матрицей с более низким числом ребер. Алгоритм был предложен в работе [Dempster (1972)] и называется алгоритмом выбора ковариаций или алгоритмом Демпстера, он подробно представлен в разделе 3 данной статьи.

2.2. Алгоритмы выбора модели

Графовые модели задаются только прямыми взаимодействиями. В нашем случае задание нулевого прямого взаимодействия между двумя элементами соответствует приравниванию нулю соответствующего элемента в обратной корреляционной матрице.

В литературе предлагаются три алгоритма выбора модели:

- методы пошагового поиска;
- EH-процедура;
- выбор с применением информационных критериев.

Эти алгоритмы кратко представлены ниже.

Пошаговый (обратный или прямой) выбор

Это возрастающая процедура поиска. Начиная с некоторой начальной модели, ребра добавляются или удаляются последовательно до тех пор, пока не удовлетворяется некоторый критерий. На каждом шаге проводится тест значимости, чтобы решить вопрос о включении или исключении предлагаемых к рассмотрению ребер. При прямом выборе начинаем с пустой модели и на каждом шаге добавляем самые значимые ребра. Напротив, в обратной процедуре выбора начинаем с полной модели и на каждом шаге удаляем наименее значимые ребра.

Оба подхода часто приводят к весьма похожим результатам и имеют свои достоинства и недостатки.

Прямой выбор начинается с пустой модели и продолжается путем перебора моделей, не согласующихся с данными, тогда как обратный метод начинается с полной модели, согласующейся с данными, который продолжается затем шаг за шагом путем перебора моделей, согласующихся с данными. С другой стороны, как отмечалось Эдвардсом [Edwards (2000)], прямой выбор имеет меньше проблем, связанных с существованием оценок максимального правдоподобия и точностью асимптотических распределений.

Алгоритм Демпстера принадлежит к классу пошаговых алгоритмов выбора, и, как показано в параграфе 3.1, его сложность равна $\mathbf{O}(p^7)$.

EH-процедура

В EH-процедуре (названной в честь авторов работы [Edwards, Havránek (1987)]) применяется более сложный алгоритм поиска, позволяющий глобально подходить к самому перебору моделей и с помощью которого выбирается целый ряд моделей. Алгоритм основан на принципе согласованности в том смысле, что если отклоняется какая-либо одна модель, то также отклоняются все ее подмодели. Аналогично, если модель принимается, то все модели, которые ее содержат, также принимаются [Gabriel (1969)].

На любой стадии в процедуре поиска имеем три непересекающихся множества: множество всех *слабо принятых подмоделей*, которые содержат принятые модели как подмодели; множество *слабо отклоненных моделей*, которые являются подмоделями одной или более отклоненных моделей и поэтому могут также рассматриваться как не согласующиеся с данными; и множество всех других, еще неклассифицированных моделей.

Следующий шаг состоит из тестирования либо минимальной, либо максимальной модели в множестве еще неклассифицированных моделей. Как только эта модель отклонена или принята, списки принятых и отклоненных моделей обновляются, и процесс повторяется до тех пор, пока не будут классифицированы все модели.

Детальный анализ EH-процедуры показывает, что можно рассматривать ее как две пошаговые процедуры (прямую и обратную), которые сходятся друг к другу. Таким образом, EH-процедура имеет, по крайней мере, такую же сложность, что и стандартная пошаговая процедура.

В EH-процедуре применяются тесты на полное качество приближения данных моделью, а не тестирование между последовательными моделями. Модели, выбранные в соответствии с EH-процедурой, как правило, будут более простыми, чем выбранные пошаговыми методами. Тем не менее известно, что полные тесты качества приближения данных моделью

часто являются менее надежными, чем тест, основанный на разностях отклонений между последовательными моделями.

Выбор с применением информационных критериев

Акаике [Akaike (1974)] предложил информационный критерий, полученный по принципу максимизации информации. Согласно этому критерию выбирается модель с наименьшим значением $-2\log(L)+2p$, где L — максимизированная функция правдоподобия модели, а p — размерность (число свободных параметров) модели.

Байесовский информационный критерий Шварца [Schwarz (1978)] асимптотически соответствует выбору модели с наибольшей апостериорной вероятностью. С помощью этого критерия проводится поиск модели с наименьшим значением $-2\log(L)+2\sqrt{n}$, где n — число наблюдений. В работе [Schwarz (1978)] также представлены и некоторые другие информационные критерии.

Выбор, основанный на оптимизации одного из информационных критериев, концептуально прост, но в вычислительном отношении может быть невыполним для моделей с большим количеством переменных.

2.3. Модели с «новым алгоритмом выбора»

Большое количество переменных (около 20–40 переменных) — это обычное число переменных в экономических исследованиях. В качестве начальной точки алгоритма выбора ковариаций Демпстера предлагаются классические алгоритмы выбора, использующие древообразные структуры зависимостей (или более точно, первые ребра древообразных структур зависимостей). Такое решение подробно представлено во второй части статьи.

Оставшаяся часть статьи посвящена подробному изложению алгоритма Демпстера.

3. Алгоритм выбора ковариаций Демпстера

3.1. Описание алгоритма

Как упомянуто выше, Демпстер [Dempster (1972)] предложил алгоритм выбора ковариаций, являющийся итерационным прямым алгоритмом выбора для матричной аппроксимации.

Три основных шага алгоритма представлены ниже, и они рассматриваются более детально в последующих разделах. Будем рассматривать только верхний треугольник ковариационной матрицы и элементы матрицы, связанные со взвешенными ребрами графа.

Шаг инициализации. Все переменные рассматриваются как некоррелированные. Ковариационная матрица является диагональной. Для данного распределения вычисляется логарифмическая функция правдоподобия.

Шаг I. Среди всех элементов исходной ковариационной матрицы выбирается тот еще не добавленный элемент, для которого повторно оцененная ковариационная матрица приводит к максимальному значению логарифма правдоподобия.

Шаг II. Если этот новый элемент вносит достаточный вклад[5], то он добавляется и производится возврат к шагу I. В противном случае процесс останавливается.

[5] Тестируется на значимость разность между новым и старым значением логарифмической функции правдоподобия.

Определения

Плотность многомерного вектора наблюдений $x = (x_1, x_2, \dots, x_p)'$ с нулевым математическим ожиданием, принадлежащего экспоненциальному семейству распределений, можно выразить в следующем виде:

$$f(\mathbf{x}, \mathbf{\Phi}) = \exp(\varphi_0 + t_0(\mathbf{x}) + \varphi_1 t_1(\mathbf{x}) + \dots + \varphi_r t_r(\mathbf{x})), \tag{8}$$

где $\int f(\mathbf{x}, \mathbf{\Phi}) d\mathbf{x} = 1$.

Величины $\varphi_1, \dots, \varphi_r$ называют естественными параметрами семейства распределений[6].

В частности, рассматривается плотность семейства многомерных нормальных распределений

$$f(\mathbf{x}, \Sigma) = \left(\frac{1}{2\pi}\right)^{\frac{p}{2}} \left(\frac{1}{\det\Sigma}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2}\mathbf{x}'\Sigma^{-1}\mathbf{x}'\right), \tag{9}$$

где $\Sigma = (\sigma_{ij})$, а $\Sigma^{-1} = \{\sigma^{ij}\}$ — ее обратная матрица.

Особый интерес представляет обратная ковариационная матрица, поскольку, как упоминалось ранее, две переменные — частично независимые (независимые при фиксированных всех остальных переменных), если и только если соответствующий элемент σ^{ij} обратной ковариационной матрицы равен нулю.

Выражение (9) можно переписать в виде (8), где:

$$r = p(p+1)/2,$$
$$\varphi_1 = \sigma^{11}, \varphi_2 = \sigma^{12}, \dots, \varphi_p = \sigma^{1p}, \varphi_{p+1} = \sigma^{22}, \varphi_{p+2} = \sigma^{23}, \dots, \varphi_{2p-1} = \sigma^{2p}, \dots, \varphi_r = \sigma^{pp}, \tag{10}$$

$$t_1(\mathbf{x}) = -\frac{1}{2}x_1^2,\ t_2(\mathbf{x}) = -x_1x_2, \dots, t_r(\mathbf{x}) = -\frac{1}{2}x_p^2, \tag{11}$$

$$t_0(\mathbf{x}) = 0, \text{ и } \varphi_0 = -\frac{p}{2}\log 2\pi - \frac{1}{2}\log(\det\Sigma), \tag{12}$$

где p — число переменных.

Отметим, что σ^{ij} являются естественными параметрами модели (8), и это представление наводит на мысль, что свертку параметров можно разумно провести, задавая равными нулю некоторые из элементов σ^{ij}.

Это — фундаментальная идея подхода Демпстера, и вообще графовые гауссовские модели определяются путем задания равными нулю определенных элементов обратной ковариационной матрицы, а следовательно, и соответствующих частных коэффициентов корреляции.

Кроме естественных параметров $\varphi_1, \varphi_2, \dots, \varphi_r$, удобно рассматривать моментные параметры $\Theta = (\theta_1, \theta_2, \dots, \theta_r)'$, определяющиеся соотношением:

$$\theta_i = E[t_i(x)] = \int t_i(x) f(\mathbf{x}, \mathbf{\Phi}) d\mathbf{x}.$$

Для нормального распределения имеем

$$\Theta = \left(-\frac{1}{2}\sigma_{11}, -\sigma_{12}, \dots, -\sigma_{1p}, -\frac{1}{2}\sigma_{22}, -\sigma_{23}, \dots, -\sigma_{2p}, \dots, -\frac{1}{2}\sigma_{pp}\right)'.$$

[6] Эти параметры полностью характеризуют распределение вплоть до линейного преобразования.

Таким образом, Θ — функция элементов ковариационной матрицы Σ, причем дисперсии умножаются на множитель $-\frac{1}{2}$ и ковариации — на множитель –1.

Определяем $r \times r$ ковариационную матрицу Γ с элементами γ_{ij},

где $$\gamma_{ij} = \mathrm{cov}(t_i(x), t_j(x)) = \int (t_i(x) - \theta_i)(t_j(x) - \theta_j) f(x, \Phi) d\mathbf{x}. \quad (13)$$

Матрица Γ по построению является положительно определенной матрицей. Ковариацию γ_{ij} для $t_i(x)$ и $t_j(x)$ можно получить, применяя стандартную формулу

$$\mathrm{cov}(x_k x_l, x_m x_n) = c_1 \sigma_{km} \sigma_{ln} + c_2 \sigma_{kn} \sigma_{lm}, \quad (14)$$

в которой коэффициенты $c_1, c_2 \in \left\{-\frac{1}{2}, 1\right\}$ соответствуют различным комбинациям σ_{kl} и σ_{mn}.

Демпстер показал, что матрица Γ дает частные производные параметров θ по параметрам φ

$$\gamma_{ij} = \partial\theta_i / \partial\varphi_j, \quad (15)$$

и вследствие симметрии матрицы Γ имеем

$$\partial\theta_i / \partial\varphi_j = \partial\theta_j / \partial\varphi_i. \quad (16)$$

Эти два свойства матрицы Γ служат основанием для итерационной процедуры, описываемой ниже.

Итерационная процедура

Рассматриваем только верхнюю треугольную часть ковариационной матрицы. На каждом шаге разбиваем множество ***F*** всех пар (i, j) на два подмножества A и B, здесь $1 \le i \le j \le p$. Подмножество A составлено из всех пар (i, j), где $i = j$ (диагональные элементы) плюс все пары (i, j), соответствующие элементам, включенным в модель, а B — подмножество всех оставшихся пар (i, j).

Такое разделение естественно приводит к разбиению $\Phi = (\Phi_A, \Phi_B)'$ и $\Theta = (\Theta_A, \Theta_B)'$. Соответственно, матрица Γ разбивается следующим образом:

$$\Gamma = \begin{bmatrix} \Gamma_{AA} & \Gamma_{AB} \\ \Gamma_{BA} & \Gamma_{BB} \end{bmatrix}.$$

Вспомним уравнения (10) и тот факт, что Φ соответствует обратной ковариационной матрице Σ^{-1}. Подбираем Θ_A установкой $\Phi_B = 0$. Таким образом, удовлетворяем определению 8 оценок $\hat{\Sigma}$ и $\hat{\Sigma}^{-1}$, представленному в следующем разделе.

В итерационной процедуре, которая строит разбиения $\Phi^{(i)} = (\Phi_A^{(i)}, \Phi_B^{(i)})$, сохраняем $\Phi_B^{(i)} = \Phi_B = 0$ так, чтобы с изменением i изменялась только $\Phi_A^{(i)}$ таким образом, чтобы для $\Theta^{(i)} = (\Theta_A^{(i)}, \Theta_B^{(i)})$ величины $\Theta_A^{(i)}$ сходились к желаемой Θ_A.

Следующий алгоритм можно рассматривать как приложение метода Ньютона для решения неявных уравнений. Если Φ_B — фиксирована, то Θ_A можно рассматривать как функцию $\Theta_A^{(i)} = \Theta_A(\Phi_A^{(i)})$ на шаге i. Используя свойство (15), разлагаем Θ_A в ряд Тейлора в окрестности Φ_A:

$$\Theta_A = \Theta_A^{(i)} + \Gamma_{AA}^{(i)}(\Phi_A - \Phi_A^{(i)}) + \cdots$$

Исключив члены более высокого порядка и заменив Φ_A на $\Phi_A^{(i+1)}$, имеем:

$$\Theta_A = \Theta_A^{(i)} + \Gamma_{AA}^{(i)}(\Phi_A^{(i+1)} - \Phi_A^{(i)}) + \cdots.$$

Решая линейную систему

$$\Gamma_{AA}^{(i)}(\underbrace{\Phi_A^{(i+1)} - \Phi_A^{(i)}}_{s}) = \Theta_A - \Theta_A^{(i)}, \tag{17}$$

получим

$$\Phi_A^{(i+1)} = \Phi_A^{(i)} + s. \tag{18}$$

Правило решения, определяющее выбор следующего недиагонального элемента для включения в *A,* состоит в том, чтобы выполнить итерацию соответствующей процедуры для всех кандидатов и выбрать кандидата, который вносит наибольший вклад в логарифмическую функцию правдоподобия.

Свойства оценки

Пусть *S* — выборочная ковариационная матрица

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})'. \tag{19}$$

Далее продолжаем изложение уже для наблюдаемой матрицы *S* (вместо матрицы Σ). Для данного разбиения (A, B) матрицы $\hat{\Sigma}$ и $\hat{\Sigma}^{-1}$ вычисляются, чтобы удовлетворить определению 8, даваемому ниже.

Определение 8 ($\hat{\Sigma}$ и $\hat{\Sigma}^{-1}$).
$\hat{\Sigma}$ *определяется так, чтобы быть положительно определенной симметрической матрицей, такой, что*

$$\hat{\sigma}_{ij} = s_{ij}, \ \textit{если} \ (i,j) \in A$$

и

$$\hat{\sigma}^{ij} = 0, \ \ \textit{если} \ (i,j) \in B.$$

Оценки $\hat{\Sigma}$ и $\hat{\Sigma}^{-1}$ обладают следующими свойствами.

Существование и единственность. Если имеется какая-либо положительно определенная симметрическая матрица, которая согласуется с матрицей *S* в позициях (i, j) множества пар *A*, тогда существует одна и только одна матрица $\hat{\Sigma}$ с дополнительным свойством, что элементы $\hat{\Sigma}^{-1}$ равны нулю в позициях *B*.

Оценивание методом максимального правдоподобия (*МП*). Среди всех нормальных моделей с элементами Σ^{-1}, равными нулю в позициях B, оценка $\hat{\Sigma}$ является МП-оценкой Σ.

Правило остановки

Обозначим через μ вектор средних значений распределения. Далее рассмотрим условные логарифмические функции правдоподобия $l(\hat{\Sigma}; \Sigma|\mu)$ относительно данного μ, которые можно вычислить с помощью любого из двух вариантов, основываясь на:

вариант 1. Только на оцененной ковариационной матрице

$$l(\hat{\Sigma}|\mu) = -\frac{p}{2}\log(2\pi) - \frac{1}{2}\log(\det\hat{\Sigma}) - \frac{1}{2}\sum_{i,j}\hat{\sigma}_{ij}\hat{\sigma}^{ij} = -\frac{p}{2}\log(2\pi) - \frac{1}{2}\log(\det\hat{\Sigma}) - \frac{p}{2}, \quad (20)$$

где p — число параметров;

вариант 2. На начальной и оцененной ковариационных матрицах

$$l(\hat{\Sigma};\Sigma|\mu) = -\frac{p}{2}\log(2\pi) - \frac{1}{2}\log(\det\Sigma) - \frac{1}{2}\sum_{i,j}\hat{\sigma}_{ij}\sigma^{ij}, \quad (21)$$

где $\hat{\sigma}_{ij}$ — элементы матрицы $\hat{\Sigma}$.

Первый вариант предложен в самой статье Демпстера, а второй рекомендован Л. Д. Мешалкиным[7]. Основываясь на результатах проведенного сравнительного анализа, использовался первый вариант вычисления логарифмической функции правдоподобия.

Определение 9. *Назовем δ_{k+1} вкладом оцененной ковариационной матрицы $\hat{\Sigma}$ в аппроксимацию ковариационной матрицы Σ при добавлении параметра $k+1$:*

$$\delta_{k+1} = 2n\left|l(\hat{\Sigma}_{k+1}) - l(\Sigma_{k+1})\right|, \quad (22)$$

где n — число наблюдений.

Точных тестов значимости не имеется, но существует несколько приближенных тестов. В частности, вклад δ_{k+1} можно рассматривать как χ^2 переменную с 1-ой степенью свободы.

Согласно методу Бонферрони [Miller (1981)], предлагается скорректировать уровень значимости α на число параметров, которые еще не содержатся в модели. Предполагаем, что переменная $k+1$ добавляется, если вклад δ_{k+1} значим на уровне $\alpha/(p(p-1)/2-k)$.

В следующем разделе представлен псевдокод алгоритма.

3.2. Псевдокод

Определения

p	— число переменных;
n	— число наблюдений;
$V = (X, U)$	— построенный граф;
$S = s_{ij}$	— выборочная корреляционная матрица;
$\hat{\Sigma} = \hat{\sigma}_{ij}$	— оценка корреляционной матрицы;
$\hat{\Sigma}^{-1} = \hat{\sigma}^{ij}$	— оценка обратной корреляционной матрицы;
$A = (i, j)$	— выбранные ребра $(i, j)\|1 \le i \le j \le p$ и $\hat{\sigma}_{ij} = s_{ij}$;
$B = (i, j)$	— $(i, j)\|1 \le i \le j \le p$ и $\hat{\sigma}^{ij} = 0$;
	$A \cup B = (i, j)\|1 \le i \le j \le p, A \cap B =$ Ш;
l	— логарифмическая функция правдоподобия $\hat{\Sigma}$;
$g_i = l_i - l_{i-1}$	— вклад в логарифмическую функцию правдоподобия;
Φ_A, Θ_A и Θ_B	— векторы длины $\#A$ и $\#B$;
Γ_{AA}	— матрица размера $\#A \times \#A$.

[7] Частное сообщение.

Алгоритм 1. Выбор ковариаций.

1. $A_0 = (i, j) | i = 1,2,\dots, p$
2. $X = 1, 2, \dots, p, U = \varnothing$
3. $\hat{\Sigma} = I$
4. Вычислить l_0 и инициализировать $g_1 = \infty$
5. **while** g_1 значимо **do**
6. $B_0 = (i, j) | i < j$ и $(i, j) \notin A_0, g_0 = 0$
7. **for** $(i, j) \in B_0$ **do**
8. $A = A_0 \cup (i, j), B = B_0 \setminus (i, j)$
9. Вычислить $\hat{\Sigma}_1$
10. Вычислить l_1
11. $g_1 = l_1 - l_0$
12. **if** $g_1 > g_0$ **then**
13. Выбрать (i, j)
14. **end if**
15. $l_0 = l_1$, $g_0 = g_1$
16. **end for**
17. $U = U \cup (i, j), \hat{\Sigma}_0 = \hat{\Sigma}_1$
18. $A_0 = A, B_0 = B$
19. **end while**

Следующий алгоритм детализирует оператор 9 алгоритма выбора ковариаций.

Алгоритм 2. Построение $\hat{\Sigma}_1$.

1. Вычислить $\hat{\Sigma}_0^{-1}$
2. $\Phi_B = [0,\dots,0]$
3. $\Theta_A = \theta_{(i,j)}$,

где $$\theta_{(i,j)} = \begin{cases} -\frac{1}{2}, & \text{если } i = j, \\ -s_{ij}, & \text{если } i \neq j. \end{cases}$$

4. Вычислить Γ_{AA} и Γ_{AA}^{-1}
5. Инициализировать $\Delta = \infty, \xi = 0{,}0001$
6. **while** $\Delta > \xi$ **do**
7. $\Phi_A^0 = \varphi_{(i,j)}^0 = \hat{\sigma}_0^{ij}$
8. $\Theta_A^0 = \theta_{(i,j)}^0$,

где $$\theta_{(i,j)}^0 = \begin{cases} -\frac{1}{2}\hat{\sigma}_{ij}, & \text{если } i = j, \\ -\hat{\sigma}_{ij}, & \text{если } i \neq j. \end{cases}$$

9. Решить $\Gamma_{AA} s = (\Theta_A - \Theta_A^0)$
10. $\Phi_A^1 = \Phi_A^0 + s$
11. Вычислить $\hat{\Sigma}_1^{-1}$ и $\hat{\Sigma}_1$
12. $\hat{\Sigma}_0 = \hat{\Sigma}_1, \hat{\Sigma}_0^{-1} = \hat{\Sigma}_1^{-1}$
13. $\Delta = ||\Phi_A^1||^2 - ||\Phi_A^0||^2$
14. **end while**

Массив $\Gamma_{AA} = \gamma_{(i,j),(k,l)}$ в операторе 4 алгоритма 2 строится в виде:

$$\gamma_{(i,j),(k,l)} = \begin{cases} -(\hat{\sigma}_{ik}\hat{\sigma}_{jl} + \hat{\sigma}_{il}\hat{\sigma}_{jk}), & \text{если } i \neq j,\ k \neq l, \\ -\hat{\sigma}_{ik}\hat{\sigma}_{jl}, & \text{если } i = j,\ k \neq l, \\ -\hat{\sigma}_{ik}\hat{\sigma}_{jk}, & \text{если } i \neq j,\ k = l, \\ -\frac{1}{2}\hat{\sigma}_{ik}^2, & \text{если } i = j,\ k = l. \end{cases}$$

Матрица $\hat{\Sigma}_1^{-1}$ строится из Φ_A^1 и Φ_B.

И наконец, представляем вычисление числа операций (сложность) алгоритма 1. Три оператора вносят наибольший вклад в сложность этого алгоритма:

- Оператор 5 **while** выполняется самое большее $p(p-1)/2$ раз.
- Оператор 7 **for** выполняется самое большее $p(p-1)/2$ раз.
- Оператор 9, решение линейной системы уравнений имеет сложность $\mathbf{O}(p^3)$.

Произведение операций, включенных в эти операторы, дает полную сложность алгоритма, равную $\mathbf{O}(p^7)$.

Перейдем к примеру.

3.3. Численный пример

Рассмотрим корреляционную матрицу S из примера Демпстера с числом переменных $p = 6$ и числом наблюдений $n = 72$, определенную равенством (2).

Инициализация: начинаем с подмножества B_0, составленного из всех недиагональных элементов, и подмножества A_0, составленного из всех диагональных элементов: $A_0 = \{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)\}$.

$$\hat{\Sigma} = \hat{\Sigma}^{-1} = \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix}.$$

Итерация 1: просматриваем элементы для добавления, начиная с элемента (1, 2): $s_{12} = 0{,}3966$.

Устанавливаем $A = A_0 \cup \{(1, 2)\}$ и $B = B_0 \setminus \{(1, 2)\}$.

Тогда

$$\Theta_A^{(0)} = \left(-\frac{1}{2},\ 0,\ -\frac{1}{2},\ -\frac{1}{2},\ -\frac{1}{2},\ -\frac{1}{2},\ -\frac{1}{2}\right)',$$

$$\Theta_A = \left(-\frac{1}{2},\ -0{,}3966,\ -\frac{1}{2},\ -\frac{1}{2},\ -\frac{1}{2},\ -\frac{1}{2},\ -\frac{1}{2}\right)',$$

$$\Phi_A^{(0)} = (\ 1,\ 0,\ 1,\ 1,\ 1,\ 1,\ 1\ )',$$

и $\Phi_B^{(0)}$ является нулевым вектором-столбцом длины 14.

Используя формулы (11), (13) и (14), вычисляем γ_{ij}. Например,

$$\gamma_{12} = \gamma_{(1,1),(1,2)} = -\hat{\sigma}_{11}\hat{\sigma}_{22},$$

и поэтому
$$\hat{\Gamma}_{AA}^{(0)} = \begin{pmatrix} 0{,}5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & \cdots & & & 0 \\ 0 & 0 & 0{,}5 & 0 & \cdots & & 0 \\ 0 & \cdots & 0 & 0{,}5 & & \cdots & 0 \\ 0 & & \cdots & 0 & 0{,}5 & 0 & 0 \\ 0 & & & \cdots & 0 & 0{,}5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0{,}5 \end{pmatrix}.$$

Используя формулы (17) и (18), вычисляем $\Phi_A^{(1)}$:

$$\Phi_A^{(1)} = \Phi_A^{(0)} + \hat{\Gamma}_{AA}^{(0)-1} \cdot (\Theta_A - \Theta_A^{(0)}) =$$

$$= \begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \begin{pmatrix} 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & \cdots & & 0 & 0 \\ 0 & 0 & 2 & 0 & \cdots & & 0 \\ 0 & \cdots & 0 & 2 & 0 & \cdots & 0 \\ 0 & & \cdots & 0 & 2 & 0 & 0 \\ 0 & & & \cdots & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ -0{,}3966 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Далее объединяем $\Phi_A^{(1)}$ и $\Phi_B^{(0)}$, чтобы получить $\hat{\Sigma}^{-1}$ и, таким образом, $\hat{\Sigma}$. Далее $\hat{\Sigma}$ может быть разбито на $\Theta_A^{(1)}$ и $\Theta_B^{(1)}$. Вектор $\Phi_A^{(2)}$ вычисляется по тем же формулам (17) и (18) и так далее.

На этом шаге максимальный вклад в аппроксимацию ковариационной матрицы $\delta_0 = l_1 - l_0 = 17{,}72$ достигается для элемента (4, 5).

Мы устанавливаем $A = A_0$ и $B = B_0$.

Итерационная процедура сходится к

$$\hat{\Sigma}^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ & 1 & 0 & 0 & 0 & 0 & 0 \\ & & 1 & 0 & 0 & 0 & 0 \\ & & & 1{,}2791 & 0{,}5975 & 0 & 0 \\ & & & & 1{,}2791 & 0 & 0 \\ & & & & & 1 & 0 \\ & & & & & & 1 \end{pmatrix}$$

и
$$\hat{\Sigma} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ & 1 & 0 & 0 & 0 & 0 & 0 \\ & & 1 & 0 & 0 & 0 & 0 \\ & & & 1 & -0{,}4671 & 0 & 0 \\ & & & & 1 & 0 & 0 \\ & & & & & 1 & 0 \\ & & & & & & 1 \end{pmatrix}.$$

Итерация 2: Снова пересматриваем элементы, чтобы провести добавление, начиная, как и прежде, с элемента (1, 2): $s_{12} = 0{,}3966$.

Имеем

$$\Theta_A^{(0)} = \left(-\frac{1}{2},\ 0,\ -\frac{1}{2},\ -\frac{1}{2},\ -0{,}4671,\ -\frac{1}{2},\ -\frac{1}{2}\right)',$$

$$\Theta_A = \left(-\frac{1}{2},\ -0{,}3966,\ -\frac{1}{2},\ -\frac{1}{2},\ -0{,}4671,\ -\frac{1}{2},\ -\frac{1}{2}\right)',$$

$$\Theta_A^{(0)} = (\ 1,\ 0,\ 1,\ 1{,}2791,\ 0{,}5975,\ 1{,}2791,\ 1\ )',$$

и $\Phi_B^{(0)}$ является нулевым вектором-столбцом длины 13.

Продолжаем действовать так же, как и на предыдущем шаге. Максимальный вклад $\delta_1 = 17{,}72$ достигается для элемента (1, 5). Добавляем этот элемент к A, таким образом $A = \{(1, 1), (1, 5), (2, 2), (3, 3), (4, 4), (4, 5), (5, 5), (6, 6)\}$. Результат итеративной процедуры есть

$$\hat{\Sigma} = \begin{pmatrix} 1 & 0 & 0 & 0{,}2163 & -0{,}4632 & 0 \\ & 1 & 0 & 0 & 0 & 0 \\ & & 1 & 0 & 0 & 0 \\ & & & 1 & -0{,}4632 & 0 \\ & & & & 1 & 0 \\ & & & & & 1 \end{pmatrix}.$$

Итерация 3: $\delta_2 = 17{,}39$.

…

Итерация 6: $A = \{(1, 1), (1, 2), (1, 3), (1, 5), (2, 2), (3, 3), (3, 6), (4, 4), (4, 5), (5, 5), (5, 6), (6, 6)\}$, и $B = \{(1, 4), (1, 6), (2, 3), (2, 4), (2, 5), (2, 6), (3, 4), (3, 5), (4, 6)\}$.

Получаем

$$\hat{\Sigma}^{-1} = \begin{pmatrix} 1{,}6672 & -0{,}4706 & -0{,}4929 & 0 & -0{,}6451 & 0 \\ & 1{,}1866 & 0 & 0 & 0 & 0 \\ & & 1{,}2624 & 0 & 0 & 0{,}3395 \\ & & & 1{,}2790 & 0{,}5974 & 0 \\ & & & & 1{,}7564 & 0{,}4884 \\ & & & & & 1{,}2592 \end{pmatrix}$$

и

$$\hat{\Sigma} = \begin{pmatrix} 1{,}0000 & 0{,}3966 & 0{,}3688 & 0{,}2163 & -0{,}4632 & 0{,}0802 \\ & 1{,}0000 & 0{,}1463 & 0{,}0858 & -0{,}1837 & 0{,}0318 \\ & & 1{,}0000 & 0{,}0385 & -0{,}0825 & -0{,}2376 \\ & & & 1{,}0000 & 0{,}4671 & 0{,}1708 \\ & & & & 1{,}0000 & -0{,}3656 \\ & & & & & 1{,}0000 \end{pmatrix}.$$

Вклад δ_6 равен 7,10. Этого недостаточно для добавления последнего предложенного элемента (3,6): s_{36}= 0,3395.

Обращаем внимание, что граф для этого примера совпадает с графом на рис. 1. Таким образом, этот граф представляет собой древообразную структуру зависимостей.

4. Заключение

В первой части статьи было предложено два метода анализа структуры переменных при помощи моделей на графах: древообразные структуры зависимостей и графовые модели — и уделено особое внимание алгоритму выбора ковариаций Демпстера. Также были приведены примеры использования моделей на графах для изучения структуры переменных и введены критерии качества представления исходной корреляционной матрицы.

Во второй части статьи для большого числа переменных предлагается модификация алгоритма Демпстера, значительно ускоряющая вычислительный процесс. Кроме того, представлена комплексная методика анализа структуры переменных, основанная на графовых моделях, и ее применение к сравнительному анализу российских регионов для различных временных периодов.

Список литературы

Айвазян С. А., *Енюков И. С.*, *Мешалкин Л. Д.* Исследование зависимостей. М.: Финансы и Статистика, 1985.

Заруцкий В. И. Классификация нормальных векторов простой структуры в пространстве большой размерности // Прикладной многомерный анализ. М.: Наука, 1978.

Заруцкий В. И. О выделении некоторых графов связей для нормальных векторов в пространстве большой размерности // Алгоритмическое и программное обеспечение прикладного статистического анализа. Серия: *Ученые записки по статистике*. М.: Академия наук, 1980.

Иберла К. Факторный анализ. М.: Статистика, 1980.

Прохорскас Р. П., *Жюжнис В. Е.*, *Мисюнене Н. Б.* Применение некоторых классификаторов для прогнозирования отдаленных эффектов инфаркта миокарда // Проблемы ишемической болезни сердца. Вильнюс: Мокслас, 1976. С. 261–267.

Ширяев А. Н. Вероятность-2. М.: Изд-во МЦНМО, 2004.

Akaike H. A new look at the statistical model identification // *IEEE Transactions in Automatic Control*. 1974. № 19.

Bartlett M. Contingency Table Interactions // *Journal of the Royal Statistical Society*. Supplement 2. 248–252. 1935.

Borgelt C., *Kruse R.* Graphical Models. Methods for Data Analysis and Mining. New York: John Wiley and Sons, 2002.

Castillo E., *Gutierrez J.*, *Hadi A.* Expert Systems and Probabilistic Network Models. New York: Springer-Verlag, 1997.

Chow C. K., *Liu C. N.* An approach to structure adaptation in pattern recognition // *IEEE Transactions on Systems Science and Cybernetics SSC*. 1966. SSC-2(2). December.

Chow C. K., *Liu C. N.* Approximating discrete probability distributions with dependence trees // *IEEE Transactions on Information Theory*. 1968. IT-14(1). May.

Chow C. K. Tree Dependence in Normal Distribution and Its Application in Pattern Recognition // *The 1970 International Symposium on Information Theory*. The Netherlands. 1970.

Cormen T. H., *Leiberson C. E.*, *Rivest R. L.* Introduction to Algorithms. MIT Press, 1990.

Darroch J., *Lauritzen S.*, *Speed T.* Markov fields and loglinear interaction models for contingency tables // *Annals of Statistics.* 1990. № 8.

Edwards D. Hierarchical Interaction Models (with discussion) // *Journal of the Royal Statistical Society.* Series B. 1990. № 52(1).

Dempster A. Covariance selection // *Biometrics*. 1972. № 28. March.

Edwards D. Introduction to Graphical Modelling. Second Ed. New York: Springer-Verlag. 2000.

Edwards D., *Havránek T.* A fast model selection procedure for large families of models // *Journal of American Statistical Association.* 1987. № 82.

Edwards D., *Kreiner S.* The analysis of contingency tables by graphical models // *Biometrika.* 1983. № 70.

Gabriel K. Simultaneous test procedures: Some theory of multiple comparisons // *Annals of Mathematical Statistics.* 1969. № 40.

Gibbs W. Elementary Principles of Statistical Mechanics. New Haven, Connecticut, USA: Yale University Press, 1902.

Jentzen F. An Introduction to Bayesian Networks. London: UCL Press, 1996.

Kruskal J. B. On the shortest spanning tree of a graph and the traveling salesman problem // *Proceedings of the American Mathematical Society.* 1956. № 2.

Lauritzen S. Graphical Models. Oxford, UK: Oxford University press, 1996.

Lauritzen S., *Wermuth N.* Graphical models for associations between variables, some of which are qualitative and some quantitative // *Annals of Statistics.* 1989. № 17.

Miller R. G. J. Simultaneous statistical inference. Second Ed. Berlin, New York: Springer-Verlag, 1981.

Schwarz G. Estimating the dimension of a model // *Annals of Statistics*. 1978. № 6.

Weinberg A. Quantitative analysis of the situation and development of Russian regions during the transition period. Thèse de Doctorat. Geneva: University of Geneva, 2007.

Whittaker J. Graphical Models in Applied Multivariate Statistics. Baffins Lane, Chichester: John Wiley & Sons Ltd, 1990.

Wright S. Correlation and Causation // *Journal of Agricultural Research.* 1921. № 20(7).